Coastal and Land Use Land Cover Area Recognition from High-Resolution Remote Sensing Images using a Novel Multimodal Attention Inception Residual Deep Network

Muhammad Attique Khan*, *Member IEEE*, Ameer Hamza, *Member IEEE*, Wardah Ibrar, LEILA JAMEL, Areej Alasiry, Mehrez Marzougui, Saru Kumari, Yunyoung Nam*, *Member IEEE*

Abstract— **As an important problem in earth observation, aerial scene classification tries to assign a specific semantic label to an aerial image. The land use land cover (LULC) classification in the aerial scene through remote sensing (RS) data is a key research area due to the important applications such as climate change, agriculture, urban structure, and water resources. However, classification from low-resolution remote sensing images causes accuracy and precision rate degradation. Deep learning (DL) models have shown significant performance in recent years for aerial scene classification. In this work, we proposed a novel end-to-end deep learning framework for LULC classification from RS images. The proposed framework is based on two novel deep learning architectures: Super-Resolution Residual Attention-Based Network (SR-RAN^5) and Multimodal Inception Attention-Based CNN (M^{2}IAN). In the first stage, a novel SR-RAN5 architecture is designed to generate high-quality RS images from the original datasets. M^{2}IAN architecture is proposed for the RS image classification in the second stage. The hyperparameters of the proposed M^{2}IAN model are selected through the Redfox optimization algorithm utilized in the training phase. After the training phase, the proposed model is tested on a test set of the selected datasets that are finally utilized for the classification. In addition, the proposed model is also interpreted through the LIME XAI technique, which localizes the important features of the input image. Extensive experiments are conducted on three publically available aerial scene datasets: NWPU, MLRSNet, and Mixed Coastal. On these datasets, the proposed model obtained improved accuracy of 91.8, 90.6, and 90.0%, respectively. Based on the results obtained, ablation studies, and comparison with existing techniques, it is observed that the proposed model achieves state-of-the-art performance. Also, it can be helpful in the real-time RS applications.**

Index Terms— **Remote sensing; high-resolution images; deep learning; optimization; interpretation.**

I. Introduction

RRecent advances in remote sensing (RS) technologies have played an essential role in resource management, such as water, land fragmentation, and habitat quality [1-3]. Also, remote sensing imagery (RSI) helps achieve goals for land use and land cover applications, such as land used for cultivation, urban structures, water bodies, dense populations, and industrialization [4-6]. These LULC images help monitor tasks, structural planning and analysis, and maintain resources [7]. RSI is crucial for processing massive data on surface situations, either land cover or land use [8, 9]. Due to the use of High-Resolution Remote Sensing (HRRS) imagery techniques, it becomes easy to gather multi-temporal and multi-source images from diverse geological locations [10, 11]. However, the diversity in acquired data introduces complex patterns like geometrical and object structure, which introduces new classification challenges for existing methods [12]. Acquiring RSI through different photographic settings introduces distortions, scale in variations, and illumination effects become part of the dataset (as noise) becomes part of the dataset [13]. Studies show that spectral and Spectral-Spatial SS features are currently used for the classification of RSI [14]. Due to enhanced spatial resolution, these features are not enough to capture the complete contextual information from HDRS images.

To handle these issues, deep learning (DL) models have gained attention because they can capture semantic information and properties of high-quality images like scene classification, object detection, tasks including image retrieval, and classification of LULC images [15, 16]. However, DL models faced two key challenges, such as insufficient and well-

Funding: This work is supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R897), Princess Nourah bint Abdulrahman University, Riyadh,Saudi Arabia. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/471/46. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176) and the Soonchunhyang University Research Fund.

Corresponding author e-mail: ynam@sch.ac.kr, : mkhan3@pmu.edu.sa).

Muhammad Attique Khan and Wardah Ibrar are with Center of AI, Prince Mohammad bin Fahd University, Al-Khobar, KSA (mkhan3@pmu.edu.sa).

Ameer Hamza is with Centre of Real Time Computer Systems, Kaunas University of Technology, Lithuania.

Areej Alasiry, Mehrez Marzougui are with College of Computer Science, King Khalid University, Abha 61413, Saudi Arabia (areej.alasiry@kku.edu.sa; mhrez@kku.edu.sa)

Leila Jamel is with department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428,Riyadh 11671, Saudi Arabia (Lmjamel@pnu.edu.sa)

Saru Kumari is with department of Mathematics, Chaudhary Charan Singh University, Meerut, India (saryusiirohi@gmail.com)

Yunyoung Nam is with Department of ICR Convergence, Soonchunhyang University, South Korea. (Corresponding author e-mail: ynam@sch.ac.kr, attique.khan@hitecuni.edu.pk).

annotated datasets [17] and a high number of learnable parameters. These challenges cause multi-class classification problems where the inter and intra-class similarity is high [18, 19]. Moreover, the low resolution images impact on the training of a deep model that later degrade the prediction performance. In addition, the deep models takes millions of data for the training and several researchers believe on the data augmentation step. The augmentation step generated some images based on the original data using traditional approached that may impact on the biasness of the model. Moreover, many techniques introduced in the literature that based on the contrast enhancement based data augmentation, DFDA algorithm, and label augmentation. In addition to this, the researchers focused on pre-trained models, transfer learning, and fusion models information for the accurate classification of RS images into a relevant class such as beach, water, river, and a few more as mentioned in Figure 1. Moreover, several classes contains similar patterns; therefore, it is a high chance of misclassification [20].

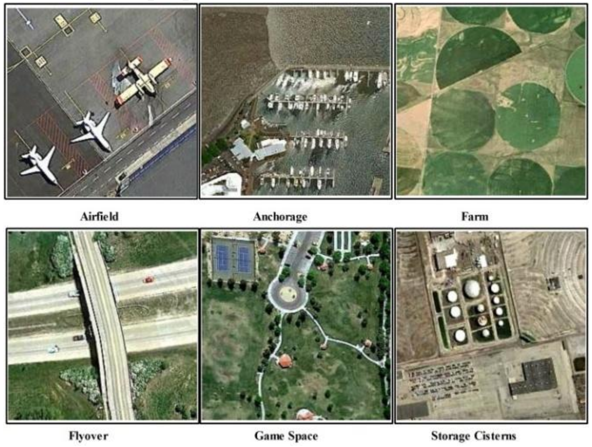


Figure 1: Sample remote sensing images collected from NWPU dataset [21]

Recent studies highlight the innovative applications of DL in the area of RS for LULC classification and urban planning [22, 23]. Fayaz et al. [24] introduced a method for classifying Land-Cover Area (ACA) using high-resolution remote sensing imagery. The significant applications of this work were urban planning, zoning agricultural land, and monitoring environmental changes. To consider these applications, an investigative study is presented that is based on the transfer learning (TL) based DCNN models training. They used Inception-V3 and DenseNet121 deep architectures and obtained an improved 92% accuracy and F1-Score on the UC-Merced-Land Use dataset. Albarakati et al. [25] introduced a novel DL architecture based on a self-attention mechanism and network-level fusion to classify LULC using RS images. They employed self-attention in designing CNN to extract important information and fuse it with CNN based on the network-level fusion approach. Moreover, they also considered the data imbalance problem and used an improved contrast enhancement technique that significantly increased the model's accuracy by 98.23%. Aljebreen et al. [26] presented a Land-Use and Land-Cover (LULC) classification technique from RS images using DL. They initially resolved the issue of noise in the original data that misleads the accurate classification. For this purpose, they used the River Formation Dynamics Algorithm (RFDA) and passed it to the Dense EfficientNet model for deep feature extraction. The extracted features were finally classified through a Softmax classifier, and a precision rate of 90.79% was obtained.

Liu et al. [27] presented a scale-aware deep reinforcement learning technique in high-resolution RS images for the Aerial Scene classification task. They consider the challenges of the DL techniques, such as dividing the images into multiple patches due to their large size. For this challenge, they design a Scale Aware Network (SAN) that utilizes images of High Spatial Resolution (HSR) and deploy a Deep Reinforcement Learning (DRL) model. Moreover, they also used a feature indexing module to identify the current patch's location. Vinaykumar et al. [28] proposed a hybrid approach using an Optimal Guidance-Whale Optimization Algorithm (OG-WOA) and deep learning networks for LULC classification. The OG-WOA technique has reduced the issue of overfitting in satellite image classification due to an improvement in the exploitation of the search process. This has been achieved through dynamic adjustment of the position of the search agents based on the best fitness value that optimizes the feature selection process. The presented OG-WOA-Bi-LSTM framework outperformed the baseline based on the improved 97.12% AID dataset classification accuracy. Moreover, the presented work obtained 96.73% accuracy on the NWPU dataset. The study shows the potential of CNN-based models for satellite image super-resolution but highlights challenges like insufficient training data and imbalanced datasets. The OG-WOA–Bi-LSTM model exhibits robust performance and reliability in satellite image classification. Xie et al. [29] introduced an automated remote sensing image scene classification approach based on incorporating Label Augmentation (LA) and an intra-class constraint mechanism. The presented approach describes the limited training sample sizes in RS datasets and the problem posed by traditional data augmentation methods that can change the content of images but keep labels. LA assigns shared labels to augmented images, representing the category information better. Moreover, Kullback-Leibler divergence is used to restrict the output distributions of images with the same scene category, reducing intra-class diversity induced by data augmentation. The presented method was tested on the UCM, AID, and NWPU datasets and obtained an average accuracy of 91.05%. Xu et al. [30] presented a Deep Feature Aggregation Framework driven by a Graph Convolutional Network named DFAGCN for scene classification from RS images. Traditional CNNs may obtain fine-classification results but fail well to realize potential context relationships in the RS image. DFAGCN solves this problem using graph-based deep learning to explore patch-to-patch correlations and intrinsic attributes. This framework uses a pre-trained VGGNet-16 model to obtain multilayer convolutional features, which are then represented as an adjacency graph. Graph Convolutional Networks reveal spatial structures and context relationships and produce refined feature embeddings. This approach combines multilayer convolutional features and fully connected features with two coefficients, α and β, for weighted concatenation further to enhance the discriminative ability of the pre-trained CNN. Fatima et al. [31] adopted a hybrid architecture (FMANet) by using both an integrated super-resolution module and a fused self-attention and inverted bottleneck CNN in a joint application for remote sensing image classification. Initially, FMANet improved the image quality by using its custom deep super-resolution (C-DSR) network, before training two feature learning networks based on two models (residual and inverted bottleneck modules) on high-resolution images. The experiments involved comparison analysis on three remote sensing datasets MLRSNet, Bijie Landslide, and Turkey Earthquake 2023 and achieved remarkable classification accuracies of 91.0%, 92.8%, and 99.4%, respectively. The model was also examined using LIME to further develop the interpretability of the process from super-resolved image to confusion matrix, however, data processing and computational limitations provide problems relative to the model complexity via depth and use of multiscale modules. Situ et al. [32] presented an attention-based deep learning framework used for the assessment of urban flood damage and risk, as well as for flood predictions and land use segmentation. The framework had two modules: the LSTM-SegNet-MSA model for estimating flood hazards, and an ES-DeepLab model to perform fine-grained land use segmentation. The LSTM-SegNet-MSA model obtained spatial (SegNet) and temporal (LSTM) feature sets that were enhanced with multi-head self-attention (MSA) and the ES-DeepLab model combined EfficientNet-Lite4 SE modules. The framework was validated against a case study in northern China where the research team accurately estimated expected annual damage (EAD) and concluded that roads and dense buildings were significant contributors to damage loss. However, although the framework was robust in estimating EAD, there are a few drawbacks with the technology presented in the framework. Their main limitation was with the use and reliance on obtaining high-quality annotated datasets. A second limitation with their framework was the prolonged computations of the attention modules for training and inference. In a study by Bhatti et al. [33] a multimodal deep learning method included a network-level fusion of a stacked residual self-attention based CNN (SRAN3) and a lightweight vision transformer (LViT-4E), with four encoders to classify high-resolution coastal area and land use and provided two approaches to complete the fusion - depth concatenation and used Bayesian optimization to tune the hyperparameters. The two models were evaluated using the EuroSAT and NWPU_RESIS45 datasets and received classification accuracies of 98.4 and 94.7% respectively. The authors noted the fused approach model performance outperformed the stand-alone models SRAN3 and LViT-4E. The multimodal approach showed promise however, the drawback was network-level fusion expands model size which could have implications in potential for real-time implementation, in addition, there are overheads connected to the training of the Bayesian optimization and deep attention modules. Ullah et al.[34] built a new wide band ensemble that incorporated Deep Snap (DS) sampling with Smooth Wavelet Convolutional Neural Networks (SWCNN), as a method to classify hyperspectral images (HSI). In this study, the researchers successfully combined spatial-spectral deep learning techniques because the researchers segmented large images into deep shots, or patch samples, while relying on multiple CNN-based learners who handled the feature extraction using discrete wavelet transforms. Because signficantly, each subsequent layer obtained images at a reduced spatial resolution. The final aspect of their classification utilized a boosting-based ensemble strategy to aggregate the predictions, across the samples, of individual models. Their method was evaluated on three open access datasets (Indian Pines, Pavia University, and Salinas Scene). As a result, they significantly outperformed similar counterparts and achieved overall accuracy scores of 98.65%, 99.52 and 99.64%, respectively, absolute scores. In addition, other potential weaknesses of the study included the demands of the computational models, especially with respect to the ensemble training stage, though overall, taking into consideration the required patch size from the employed method and tuning of multiple CNNs, the overall application of large-scale HSI could all be barriers to real-time applications.

In summary, these presented techniques focused on the dataset augmentation, improved the quality of images through traditional techniques, employed pre-trained models, and fusion at the network side. However, they faced the challenge of high number of learnable parameters, complex CNN networks that trained on the high-resolution images; earlier networks either overlook color-based traces critical to distinguishing visually similar coastal scenes, and manual hyperparameters selection. In this work, we focused on the design two different lightweight deep learning architectures for the generation of high-resolution images and accurate prediction of RS images into their relevant class. Moreover, the model is trained through automatic selected hyperparameters using Bayesian Optimization. Our key contributions in this work are as follows:

- We proposed a novel Super-Resolution Residual Attention-Based Network (SR-RAN^5) to enhance the spatial quality of remote sensing coastal areas.
- We proposed a novel multimodal Inception Attention-Based CNN (M^{2}IAN) architecture to learn the diverse and multi-scale features necessary to capture complex spatial patterns and critical transformations in coastal environments.
- We integrate color features as secondary input modality that enable the model to capture discriminative features related to chromatic deviations in diverse environments.
- Red fox optimization algorithm is implemented and selected the best hyperparameters values for the initial assignment instead of hit and trial approach.
- A detailed ablation study is conducted using the original, generated, and a mixture of original and generated data to check the model's generalizability and effectiveness.

II. Dataset and Preprocessing

In this work, two publicly available datasets are collected to evaluate the proposed model. The selected datasets are NWPU_RESISC45 and the Mixed Coastal dataset (prepared using NWPU_RESISC45, MLRSNet, SIRI, and EuroSAT).

Mixed Coastal Dataset: In this research, we acquired the relevant coastal classes from different databases such as EuroSat, SIRI, NWPU, and MLSRNet datasets for the experimental process. The collected datasets are publically available at (https://1drv.ms/u/s!AnoH-ohvSrcVbimm4dCSqGBnTS4?e=0fu0xO). The dataset has a total of 13 classes. The dataset contains 13 classes, including Anchorage, beach, harbor, harbor & port, island, lake, landslide, red sea fish, river, snow berg, swimming pool, water, and wetland; the size of each sample is 256 ×256 ×3 and nature of samples are RGB. The total number of samples in the collected dataset is 9206. A few sample images of this dataset are shown in Figure 2.

NWPU_RESISC45 Dataset: The NWPU dataset is publicly available at (https://figshare.com/articles/dataset/NWPU-RESISC45_Dataset_with_12_classes/16674166?file=30871912). It consists of 12 classes, including Airfield, anchorage, beach, dense residential, farm, flyover, forest, game space, parking space, river, sparse residential, and storage cisterns. It contains a total of 10,500 samples in all the classes. The size of each sample in the dataset is 600×600.

Both datasets have different contrast and resolution variations, which is unsuitable for the better learning of the deep learning model, as shown in Figures 1-2. Therefore, we proposed an SR-RAN5 network to generate new high-resolution samples for better network learning. All the generated samples are added to the original datasets to reduce the unbalancing problem. The whole description of the datasets is presented in Figure 3.

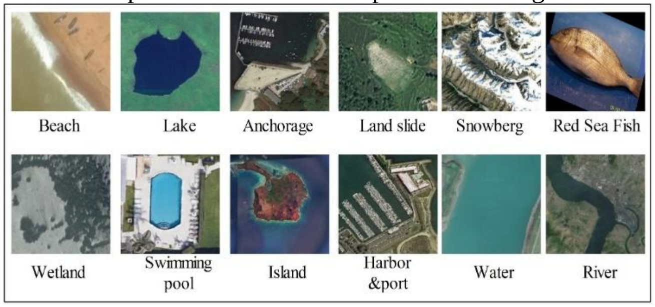
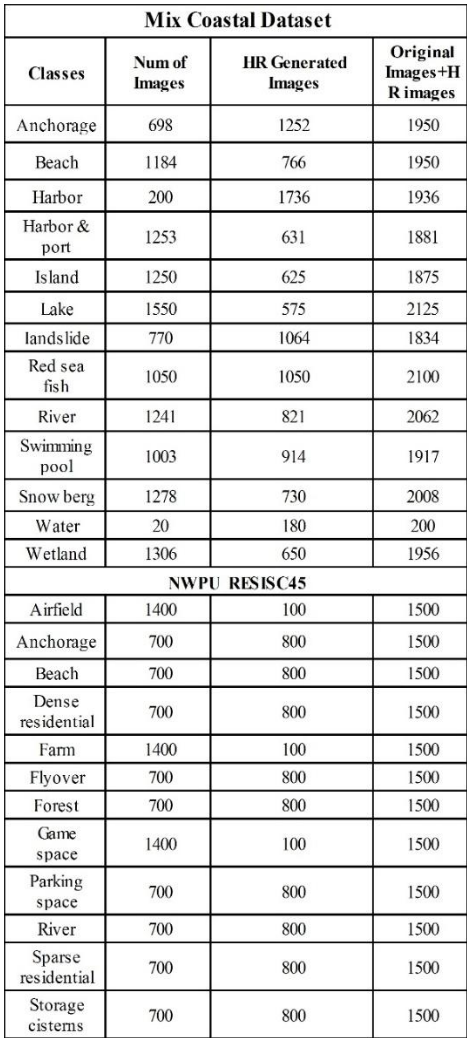


Figure 2: Few samples of the mix coastal dataset



Mix Coastal Dataset			
Classes	Num of Images	HR Generated Images	Original Images+HR images
Anchorage	698	1252	1950
Beach	1184	766	1950
Harbor	200	1736	1936
Harbor & port	1253	631	1881
Island	1250	625	1875
Lake	1550	575	2125
landslide	770	1064	1834
Red sea fish	1050	1050	2100
River	1241	821	2062
Swimming pool	1003	914	1917
Snow berg	1278	730	2008
Water	20	180	200
Wetland	1306	650	1956
NWPU RESISC45			
Airfield	1400	100	1500
Anchorage	700	800	1500
Beach	700	800	1500
Dense residential	700	800	1500
Farm	1400	100	1500
Flyover	700	800	1500
Forest	700	800	1500
Game space	1400	100	1500
Parking space	700	800	1500
River	700	800	1500
Sparse residential	700	800	1500
Storage cisterns	700	800	1500

Figure 3: Description of each class with original and generated images

III. PROPOSED SR-RAN^{5} FOR RESOLUTION ENHANCEMENT

The Very Deep Super Resolution Network (VDSR) [35] is a CNN designed for single-image super-resolution. It works by learning the mapping of low-resolution (LR) images to high-resolution (HR) images through residual learning [36]. The network usually comprises several convolutional layers, with ReLU activation and small receptor fields. Residual learning is an important feature of VDSR, simplifying training processes by allowing the network to learn residual knowledge rather than directly predicting HR images [37]. The residual-based framework enables VDSR to achieve greater convergence and reconstruction accuracy. However, the effectiveness of the VDSR is limited when rescaled to more complex datasets or captured small details in high-resolution reconstructions, mainly because there are no selective focus mechanisms on important areas [38]. To address this limitation, we proposed a 5-block residual attention-based VSDR network. The residual attention VDSR improves the standard VDSR by integrating attention mechanisms to improve the priority of features and dynamic processing capabilities. Using the attention module [39], the network can focus on the characteristic map's most important regions, ensuring better feature extraction and noise resistance. This improvement allows the network to adjust the weight of the feature maps dynamically, highlighting details in the critical areas and suppressing irrelevant information. In addition, the combination of attention modules and residual connections strengthens the learning process. It enables high-resolution images to be rebuilt more effectively, especially in areas with soft textures and high-frequency details. This cooperation allows for the Residual Attention VDSR to perform better in complex and difficult super-resolution scenarios than standard VDSR.

The proposed SR-RAN^5 accepts the input size of $227 \times 227 \times 3$. The network goal is to reconstruct the high resolution image using residual. Mathematically, it is defined as: $L_R = I_H - I_L$, where the L_R learns by the network and reconstruct the I_H so, $I_H = I_L + L_R$.

First Residual Attention Block: The initial convolutional is attached with the configuration of 3×3 kernel size, 64 kernels, and 1×1 stride. After that, an addition layer is attached to two ReLU activation layers, five convolutional with a 3×3 filter size, 64 kernels, 1×1 stride, one sigmoid, and one residual connection. The mathematical representation of the first residual attention block is defined as follows:

$$\varphi_0 = Conv(I_L; \omega_0; b_0) \quad (1)$$

$$\varphi_1 = ReLU(Conv(\varphi_0; \omega_1; b_1)) \quad (2)$$

$$\varphi_2 = Conv((\varphi_1; \omega_2; b_2)) \quad (3)$$

$$\psi_1 = ReLU\left(Conv\left(\varphi_2; \omega_{\psi^1}; b_{\psi^1}\right)\right) \quad (4)$$

$$\psi_2 = \sigma\left(Conv\left(\psi_1; \omega_{\psi^2}; b_{\psi^2}\right)\right) \quad (5)$$

$$\varphi_{skip} = \varphi_0 + \psi_2 \odot \varphi_2 \quad (6)$$

where φ_0 is the output feature map, ω_0 and b_0 is the weights and biases, σ is the sigmoid activation, ψ_2 is the attention map, and $\odot$ presented the element-wise multiplication. The other four residual attention blocks is designed based on the same phenomena and the depth size is updated. The depth sizes are 96,128, 64, and 32, respectively. After residual attention blocks, the final I_H residual image is reconstructed by attaching the final convolutional layer $L_R = Conv(\varphi_{final}; \omega_r; b_r)$. The HR image is obtained by adding the residual L_R image using $I_H = I_L + L_R$. Mean squared error is the loss function of the proposed model which is defined as $\frac{1}{T_s}\sum_{z=1}^{T_s} ||f\left(\varphi, I_{L_z}\right) - G_k||^2$. The proposed SR-RAN^5 has 682K parameters with 51 layers. The overall systematic architecture is presented in Figure 4. Moreover, a few sample generated images are shown in Figure 5. The generated high resolution samples are later replaced by the original dataset images and utilized for the training of the classification model.

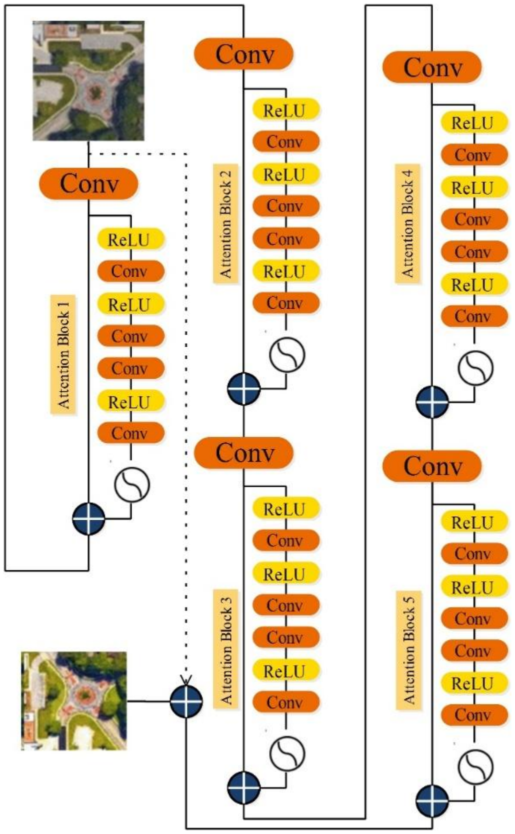


Figure 4: Architecture of Proposed SR-RAN^5 for high resolution images

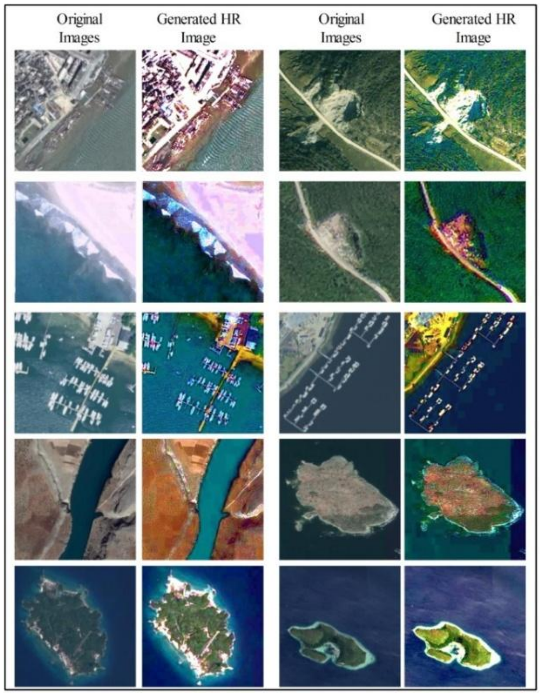


Figure 5: Visual illustration of generated high resolution images through proposed SR-RAN^5

A. Proposed Multimodal Inception Attention Network M^2IAN

A convolutional neural network (CNN) [40] is a foundation of image classification and computer vision tasks. It efficiently learns hierarchical spatial properties of input data through a convolutional, pooled, and fully connected layer. These networks can highly capture local patterns, such as edges, textures, and shapes, making them highly effective in visual data-related tasks [41]. However, CNNs inherently focus on local receptive fields and limit their ability to model long-term dependence and global contextual relationships effectively. To overcome these limitations, CNN architectures have introduced attention mechanisms. These mechanisms allow networks to dynamically prioritize the most relevant regions of the image, focusing on spatial or channel-wise importance and improving the extraction of features. Popular attention modules such as SE [42], CBAM [43], and non-local blocks enable CNNs to achieve a stronger feature representation by capturing the interdependencies of the feature map. Despite these improvements, CNN faces multimodal data processing challenges: information derived from different inputs such as images, texts, and sensor data. Multimodal CNNs [44] address this problem by integrating complementary information from various modes to create richer and more holistic representations of features. This integration usually involves feature-fusion strategies such as concatenation, weighted average, or attention-based fusion that effectively combine multimodal data. Using the strengths of each type, multimodal CNNs improve classification precision, especially in complex tasks where single-type data cannot provide sufficient discriminatory information.

In this work, we designed an inception attention module and integrated it into the network layer series. In addition, we also included color features as the second input to improve further the model's ability to analyze and distinguish complex categories such as islands, beaches, wetlands, lakes, landslides, anchors, rivers, and snowbergs. The purpose of designing the multimodal inception attention network is to learn the different and multi-scale features necessary to capture complex spatial patterns and critical differences in coastal environments. The inception module captures broad and fine spatial patterns such as coastlines, vegetation textures, and water boundaries. At the same time, the attention mechanism focuses on critical areas such as the vegetation-water interface and land landslides, effectively highlighting essential features and suppressing irrelevant information. The presence of color properties enhances the model's ability to analyze color variations and their spatial distribution, which play a crucial role in determining the types of surface areas, water boundaries, and other critical features of coastal and complex structures. This approach enables a more detailed analysis of complex coastal environments. The proposed M^{2}IAN network has two inputs: Features input and Image input.

Feature Input: This input takes colors feature vector with dimension of $N \times 1024$, where N is the number of training samples. The mathematical representation of color features is defined as follows:

Consider an input image ϕ with the dimension of $R \times W \times 3$. Where the R and W are the length and width of the image and 3 is the channel such as red, green, and blue $\phi = [\phi_r, \phi_g, \phi_b]$, the image transformed into RGB to HSV. The transformation are fallowing as: initially, computer the max, min, and difference among the max and min channel like$\phi_{min} = \min(\phi_r, \phi_g, \phi_b)$, $\phi_{max} = \max(\phi_r, \phi_g, \phi_b)$, and $\Delta d = \phi_{max} - \phi_{min}$. The Hue ϕ_h is measured using the maximum channel from $\phi_r, \phi_g, or \phi_b$. The channel ϕ_s is measured by using $\phi_s = \frac{\Delta d}{\phi_{max}}$, and the ϕ_v is the ϕ_{max}. After that, RGB is converted into lab color space using LAB transformation (ϕ_L, ϕ_A, ϕ_B). For the each color channel $(\phi_r, \phi_g, \phi_b, \phi_h, \phi_s, \phi_v, \phi_L, \phi_A, \phi_B)$, the statistical features are measured which is defined as:

$$Mean(\mu_\phi) = \frac{1}{R \times W} \sum_{k=1}^{R} \sum_{i=1}^{W} \phi_{k,i} \quad (7)$$

$$Variance(V_\phi) = \frac{1}{R \times W} \sum_{k=1}^{R} \sum_{i=1}^{W} (\phi_{k,i} - \mu_\phi)^2 \quad (8)$$

$$std(\sigma_\phi) = \sqrt{V_\phi} \quad (9)$$

$$Skewness(\varphi_\phi) = \frac{1}{R \times W \times \sigma_\phi^3} \sum_{k=1}^{R} \sum_{i=1}^{W} (\phi_{k,i} - \mu_\phi)^3 \quad (10)$$

$$Kurtosis(k_\phi) = \frac{1}{R \times W \times \sigma_\phi^4} \sum_{k=1}^{R} \sum_{i=1}^{W} (\phi_{k,i} - \mu_\phi)^4 - 3 \quad (11)$$

$$HarmonicMean(\mu_\phi^h) = \frac{R \times W}{\sum_{k=1}^{R} \sum_{i=1}^{W} \frac{1}{\phi_{k,i}}} \quad (12)$$

$$Median_\phi = Median\{\phi_{1,1}, \phi_{1,2}, \dots, \phi_{R,W}\} \quad (13)$$

$$Mode_\phi = MostFrquent\{\phi_{1,1}, \phi_{1,2}, \dots, \phi_{R,W}\} \quad (14)$$

$$C_F = [k_r, k_g, k_b, \varphi_r, \varphi_g, \varphi_b, \dots.] \quad (15)$$

Where ϕ is presented the channel from $(\phi_r, \phi_g, \phi_b, \phi_h, \phi_s, \phi_v, \phi_L, \phi_A, \phi_B)$, and $\phi_{k,i}$ is the pixel value at position(k, j). After manually extract the color features, it passes to the feature input layer and a ReLU activation function is applied on the color features to obtain the normalized feature map.

Image Input and First IAM Block: The image input takes image as input with the size of $227 \times 227 \times 3$. It attached with the first inception attention module that start with convolutional layer configured with 3×3 kernel size, 32 depth size, and 2×2 stride value. After that, two branches are attached, the first branch is inception mechanism for spatial attention and the other branch is channel attention. The inception mechanism is consist of further three branches, the first inception branch contains two convolutional layers of filter sizes 1×1 and 1×2, depth size of 16, stride value of one, batch normalization layer, and a ReLU activation layer. The second and third branch configurations are 3×1, 3×3, 1×2, 2×1 filter sizes, depth size of 16, stride value is same (one) with batch normalization and ReLU activation layers. All the three branches of inception are depth wise concatenated at the end for output weight features. A global average pooling, and two fully connected layers with 16 and 32 depth, are connected in the channel attention branch. After that, the channel and spatial attention feature maps are added using the additional layer and sigmoid function is employed on the output of the addition layer. Moreover, the multiplication layer is employed to multiply the feature map obtained from the inception attention mechanism and first convolutional layer. In the end, the resultant feature matrix is further added with the first convolutional layer.

Consider an input image $X \in \mathbb{R}^{R \times C \times D}$, where $R = C = 227$ and $D = 3$, the mathematical formulation of inception attention module is defined as follows:

$$\varphi_c = \oint(Conv(X, \omega_c) + b_c) \quad (16)$$

Spatial Attention:

$$\varphi_{1,1} = \oint(Conv(\varphi_c, \omega_{1\times1}) + b_{1\times1}) \quad (17)$$
$$\varphi_{1,2} = \oint(Conv\left(\varphi_{1,1}, \omega_{1\times2}\right) + b_{1\times2}) \quad (18)$$
$$\varphi_{3,1} = \oint(Conv(\varphi_c, \omega_{3\times1}) + b_{3\times1}) \quad (19)$$
$$\varphi_{3,3} = \oint(Conv\left(\varphi_{3,1}, \omega_{3\times3}\right) + b_{3\times3}) \quad (20)$$
$$\varphi_{3,1} = \oint(Conv(\varphi_c, \omega_{1\times2}) + b_{1\times2}) \quad (21)$$
$$\varphi_{3,1} = \oint(Conv\left(\varphi_{3,1}, \omega_{2\times1}\right) + b_{2\times1}) \quad (22)$$
$$\varphi_S = Concat(\varphi_{1,2}, \varphi_{3,3}, \varphi_{3,1}) \quad (23)$$

Channel Attention:

$$C_{GAP} = \frac{1}{R\times C}\sum_{r=1}^{R}\sum_{C=1}^{C}\varphi_c(r, c, :) \quad (24)$$
$$C_{FC1} = \sigma(w_1.C_{GAP} + b_1) \quad (25)$$
$$C_{FC2} = \sigma(w_2.C_{FC1} + b_2) \quad (26)$$
$$C_A = \sigma(C_{FC2} + \varphi_S) \quad (27)$$
$$\phi_{weight} = C_A \odot \varphi_c \quad (28)$$
$$\partial_{res} = \phi_{weight} + \varphi_c \quad (29)$$

Where $\oint$, is the ReLU activation, ω_c is the kernel sizes, b_c is the bias factor, φ_S presented the resultant feature map of spatial attention, C_A presented the resultant of channel attention feature map, ∂_{res} is the final outcome of the inception attention module, $\odot$ presented the element wise addition, and σ sigmoid function.

Stage1: Transition Layers: After the first IAM, a convolutional layer with a filter size of 1 × 1 has been added, whereas the depth size is 48, and stride value is 1 × 1. A ReLU activation layer is added after each convolutional layer. In the second convolutional layer, the filter size is 3 × 3, 64 depth value, and 1 × 1 stride value. In addition, one max pooling layer is also added of 3 × 3 pool size.

Second IAM Block: After stage 1, second IAM block is employed with same mechanism but the kernel size and depth values are updated. The employed kernel sizes and depths are (3 × 3, 64), (5 × 1, 32), (1 × 5, 22) for first inception branch, (4 × 1, 64), (1 × 4, 21) for second inception branch, and (3 × 1,64), (1 × 3, 21) for third inception branch, respectively. The depths in channel attention are 32 and 64, respectively.

Stage 2: Transition Layers: After the second IAM, two convolutional, one ReLU, and one max pooling formed with 1 × 1, 1 × 1 filter sizes, 1 × 1 pool size, 56 and 96 depths are integrated. Third IAM Block: After Transition layer, the third IAM block is also integrated with the same phenomena and the configurations of these block are (3 × 3, 96), (7 × 7, 64), and (1 × 1, 42) for first inception branch of this block, (5 × 5, 64) and (1 × 1, 43) for second branch, and (3 × 3, 64) and (1 × 1, 43) for third inception branch. The depths in channel attention are 64 and 128, respectively.

Stage 3: Transition Layers: In this stage, the same series of transition layer are attached with the setting of (1 × 1,72), (1 × 1, 114), and 3 × 3 pool size. Fourth IAN Block: The fourth IAM block is also based on the inception attention module with the formations of (3 × 3, 256), (7 × 7, 128) and (1 × 1, 85) for first inception branch, (5 × 5, 128) and (1 × 1, 85) for second inception branch, and (3 × 3, 128) and (1 × 1, 86) for third inception branch, respectively. The updated depths in the channel branches are 128 and 256, respectively.

Stage 4: Transition Layers: In the final transition block, a convolutional layer is added having 1 × 1 kernel size, 512 depth, and 1 × 1 stride value, respectively. A ReLU activation layer is added after each convolutional layer. After that, a convolutional layer of 3 × 3 kernel size, 1024 depth, and 1 × 1 stride value has been added, whereas the max pooling layer size was 3 × 3. The end of this transition layer, one global average pooling and flatten layer is added. The flatten layer converted extracted features in 1-dimensional.

Feature Concatenation and Final Layers: The color feature map and resultant feature map of the IAN is combined by using the depthwise concatenation layer. After that, a fully connected layer, a softmax, and a classification layers has been attached in the end of the network. Mathematically, these layers are defined as follows:

$$\varphi_{conc} = DepConcat(\mathrm{F}_c, \mathrm{F}_{IAN}), \text{ Where } \varphi_{conc} \in \mathbb{R}^{R\times C\times(\mathrm{F}_c+\mathrm{F}_{IAN})} \quad (30)$$
$$\psi_{FC} = W_{FC}.\varphi_{conc} + b_{fc} \quad (31)$$
$$\Phi_{prob[i]} = \frac{e^{\psi_{FC}[k]}}{\sum_i e^{\psi_{FC}[i]}}, k \in \{1,2,3, \dots, N\} \quad (32)$$
$$\Psi_{pred} = argmax_k \Phi_{prob[i]} \quad (33)$$

Where N is the total number of classes, and $\Phi_{prob[i]}$ is the probability of class k. The categorical cross entropy is employed to calculate the loss of the proposed model, which is defined as $L_{CE} = -\sum_{k=1}^{R} Q_k \log(\hat{Q}_k)$. The proposed M^{2}IAN has total 113 layer with 6.9 million parameters and 1.7 GFLOPs per forward pass . The systemic architecture is presented in Figure 6.

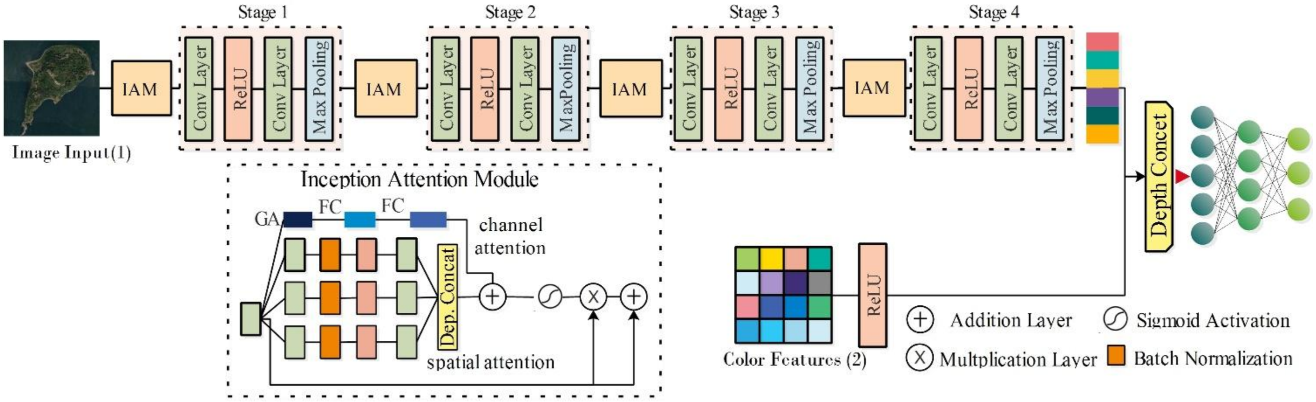


Figure 6: Architecture of proposed M^{2}IAN for Ariel Scene image classification

B. Training and Testing

The dataset is divided into a 50:50 ratio for the training and feature extraction process. The 50% of the data is employed for training, and the remaining part is used for testing. Initially, the SR-RAN^5 model is employed to obtain high-resolution images. The color features are manually extracted from the training set. The extracted color features and training set are passed to the M2IAN model for training. The proposed model required several hyperparameters for the training phase. In this work, we utilized red fox optimization to select the best hyperparameters at the initial phase. The red fox optimization is described in the below subsection.

1) Red Fox Optimization for Hyperparameters Tuning

Red Fox Optimization (RFO) algorithm [45] is natural metaheuristics that simulate the behavior of wild red fox hunting and adaptive movement. It can be used in deep learning to adjust hyperparameters by efficiently searching for the most suitable set of hyperparameters to improve the performance of the model. RFO balances exploration and use, helping to avoid local minima and find better solutions than traditional tuning methods such as grid searches and random searches. The mathematical modeling of each is step is following as:

The hyperparameters tuning involves selecting the best combinations of hyperparameters for a deep learning model. The hyperparmeter are denoted as $\varphi_H = \{\varphi_1, \varphi_2, \varphi_3, \varphi_4\}$, where $\varphi_1, \varphi_2, \varphi_3, \varphi_4$ are presented the epochs, learning rate, batch size, and activation function. For the each solution the set of hyperparameters is fallow as:

$$S_k = (\varphi_1^k, \varphi_2^k, \varphi_3^k, \varphi_4^k), k = 1,2,3,4, \ldots., N \quad (34)$$

Where N is the number of population size. In the next, the foxes initial population is randomly generated by using the following equation [45]:

$$S_k = S_{min} + ran.(S_{max} - S_{min}), \qquad ran \sim (0,1) \quad (35)$$

The loss-based fitness function is employed to evaluate the quality of hyperparameters S_k. The fitness function is defined as:

$$\psi_f(S_k) = \frac{1}{1+V(S_k)} \quad (36)$$

Where the $V(S_k)$ is the validation loss and the outcome of fitness function is denoted by the $\psi_f(S_k)$. The foxes are updating the position by using exploration and exploitation process. When the fox moves towards the prey, the position is updated by using the below equation.

$$S_k^{t+1} = S_k^t + R_1.(S_{best}^t - S_k^t) + R_2.\omega \quad (37)$$

Where the S_{best}^t is the best solution, ω is the random perturbation vector to increase the diversity, which helps to avoid the local optima, $R_1$and R_2 is the random coefficient range [0,1]. Initially, the foxes moving with high speed, therefore, to reduce the step size over the time, the moving speed become slowly, it also controls the converge to the food. Mathematically, it is defined as follows:

$$S_k^{t+1} = S_k^t + \tau(S_{best}^t - S_k^t) + R_2 \quad (38)$$

$$\tau = \tau_0.e^{-\frac{t}{T}} \quad (39)$$

Where $\tau_0$the starting hunting rate and τ is the adaptive hunting behavior, T is the maximum number of iterations and t is denoted the movement smaller. In our case, some of the hyperparameters are categorical and some are discrete. For the discrete and categorical hyperperparameters are handle using the following equations:

$$\varphi_d^k = round(\varphi_d^k\{16,32,64\}) \quad (40)$$

$$\varphi_c^k = \arg min_{v\epsilon\{ReLU,Sigmoid\}}|\varphi_c^k - v| \quad (41)$$

In the end, the algorithm is stopped by employing the stopping condition. The algorithm is stop when best fitness value does not improve in the next five iterations. The complete pseudo code is provided in Table 1. Moreover, the selected hyperparameters value are given in Table 1. In this table, the ranges are given for each hyperparameter, whereas the best values are updated for each dataset used in this work.

```
Input: $Hyperparameters \leftarrow \varphi_H = \{\varphi_1, \varphi_2, \varphi_3, \varphi_4\}$
Output: $Best\ hyperparameters \leftarrow \varphi_B = \{\varphi_{b1}, \varphi_{b2}, \varphi_{b3}, \varphi_{b4}\}$
Initialize the randomly generated population N foxes
Evaluate the fitness Function
  For each fox:
    $\psi_f(S_k) = \frac{1}{1+V(S_k)}$
Set the $S_{best}^t$
While !Stopping criteria:
    For each $S_k$:
        Randomly generate $(R_1, R_1)$
  If (Exploration phase):
            Update position $\leftarrow S_k^{t+1} = S_k^t + R_1.(S_{best}^t - S_k^t) + R_2.\omega$
        Else (prey pursuit behavior):
            Compute the $\tau \leftarrow \tau_0.e^{-\frac{t}{T}}$
            Update position $\leftarrow S_k^{t+1} = S_k^t + \tau(S_{best}^t - S_k^t) + R_2$
      Apply boundary constraints
      Handle Categorical and discrete parameters
      Update fitness value
If $S_{best}^k$ ←Better solution:
      Update $S_{best}^k$
Return the best hyperparameters
```

Table 1: Hyperparameters and their ranges for the best selection of hyperparameter using red fox tuning

Hyperparameters	Ranges
Epochs	20-70
Learning rate	0.00014-0.0015
Batch size	[8,16,32]
Activation	[ReLU, Sigmoid]

C. Testing Process

After the training process of the proposed model, the testing set is utilized first to generate the high-resolution samples using the SR-RAN5 model. After that, the generated images are employed for the feature extraction, and the features are obtained from the depth-wise concatenation activation layer. The dimensions of extracted features are N×2048. The obtained feature vector is passed to the neural network classifiers for the final classification. The detailed testing process of the proposed model is shown in Figure 7. This figure notes that the original image is passed to the first proposed network that returns a high-resolution image. The resultant image is further processed in the proposed classification model that returns output in labeled images such as Islands, Beaches, Lakes, and a few more.

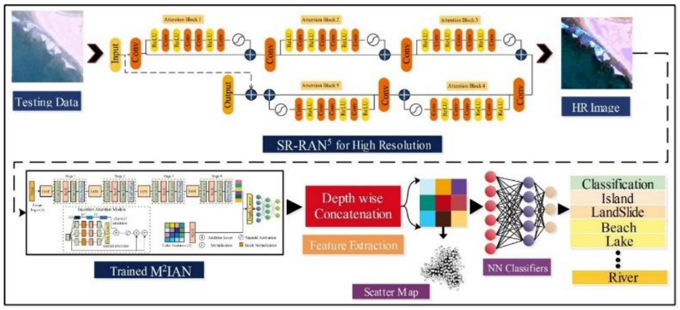


Figure 7: Systemic flow of the proposed model for testing phase

IV. Experimental Results

The experimental setup and the classification results have been presented in this section. The hyperparameters for the proposed SR-RAN^5 model are selected randomly, such as the learning rate is 0.00016, mini-batch size is 16, epochs is 200, and optimizer is Adam. For the proposed classification model M^{2}IAN, hyperparameter tuning is employed using red fox optimization. The hyperparameters and their ranges are described in Table 1. The proposed model is trained using the 5-fold cross-validation to approach better generalization. In classification, the LOOCV technique is utilized to evaluate neural network classifiers. The classifiers are assessed based on accuracy, FNR, precision, recall, f1-score, and inference time. The experiment was conducted on MATLAB R2024a using the desktop system configured with core i7 with 3.2GHz, 64GB of RAM, and an NVIDIA RTX A4500 20GB graphic card.

A. *Proposed Model Results*

The proposed model results are presented here on selected datasets such as Mix Coastal and NWPU. Dataset detail is presented under section 2. The results are presented in the form of tables, confusion matrices, and scatter plots.

1) *Classification Results on Mix Coastal Dataset*

The proposed M2IAN model is trained on the mixed coastal dataset in this experiment. The features are extracted from the test set and passed to the neural network classifiers for classification results. The classification results for this dataset are shown in Table 2. Based on this table, the WNN classifier achieves the highest accuracy of 90.0%, whereas the F1 score value is 83.789%, and the recall rate is 83.353%. The recall rate of this classifier can be proved by a confusion matrix given in Figure 8. It further supports this observation, as Harbor and Harbor&Port have significant erroneous classifications, and Harbor achieves only 55.0% accuracy due to overlapping features. However, categories such as redfish 98.3% and lake 97.6% showed high classification accuracy, indicating that models effectively distinguish well-separated feature groups.

The MNN classifier obtained the second-best accuracy of 88.5%, whereas the precision and recall rates are 82.615 and 81.415%, respectively. The other classifiers mentioned in this table, such as BNN, obtained 83.7%, and TNN, which received 82.7% accuracy, showed ineffectiveness in capturing complex patterns. In particular, WNN and MNN have lower false negative rates, which advise their reliability in minimizing misclassification. The lowest computational efficiency is 142.29 sec for the MNN classifier, and the WNN classifier has a slightly longer inference time of 151.05 sec. Figure 9 shows the distribution of feature representations across different classes. Some classes, such as redfish and landslide, clearly separate, indicating strong discriminatory characteristics, while others, such as Hart and Hart & Port, show overlap, indicating potential classification challenges. This visualization highlights the complexity of the classification task and emphasizes the importance of advanced classification for distinguishing closely related categories.

Table 2: Classification results of proposed M^{2}IAN features on NN classifiers using mix coastal dataset

Classifier	Accuracy (%)	Recall (%)	Precision (%)	FNR	Time (sec)	Pred Speed (obs/sec)	F1-Score (%)
NNN	84.7	74.938	75.1	25.06	405.1	5200	75.018
MNN	88.5	81.415	82.615	18.58	142.29	5500	82.010
WNN	**90.0**	**83.353**	**84.230**	**16.64**	**151.05**	**5300**	**83.789**
BNN	83.7	73.669	73.561	26.33	677.47	5400	73.614
TNN	82.7	71.861	71.6	28.13	906.76	5300	71.730

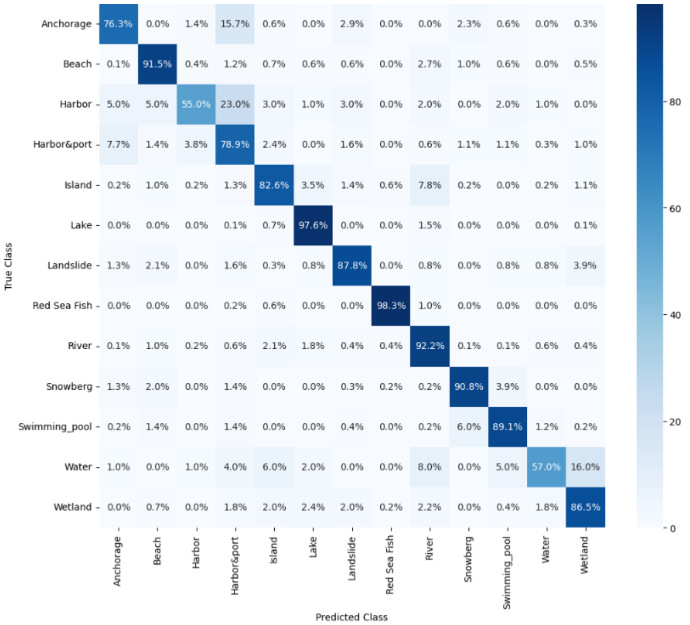


Figure 8: Confusion matric of WNN classifier on mix coastal dataset

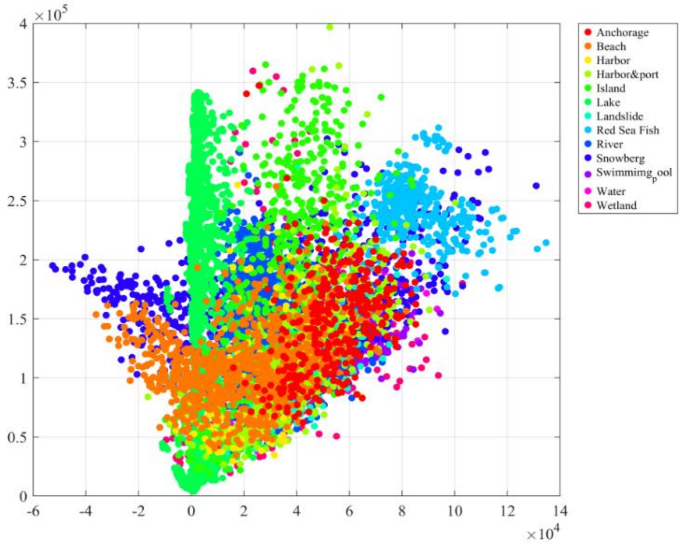


Figure 9: Distribution of features representation of each class in mixed coastal dataset

2) *Classification Results on NWPU_RESISC45*

The classification results of the proposed M2IAN model on the NWPU-RESISC dataset are summarized in Table 3 and visualized through the confusion matrix in Figure 10. The extracted features from the M2IAN model improved accuracy by 90.8%, which is superior to other classification classifiers in this table. The other listed classifiers, such as NNN, obtained 84.3% accuracy; MNN, 88.7%; BNN, 89.2%; and TNN, 91.4% accuracy, respectively. The WNN classifier showed a recall rate of 91.625%, a precision rate of 91.57%, and an F1 score of 91.59%, respectively. These high scores highlight the firm balance between determining true positives and preventing false negatives in the classifier. Compared to TNN, WNN has a slightly higher precision and recall while maintaining a competitive inference speed of 8300 obs/sec. The confusion matrix of the WNN classifier is shown in Figure 10. In this figure, it is noted that the WNN classifier shows high classification accuracy in most of the classes and is remarkable for recognizing “forest” 97.3% and “airfield” 95.2%, respectively. However, visually similar classes, such as Beach and Deep Residential, were misclassified, resulting in 4.1% and 3.5%, respectively. Overall, the proposed model obtained improved performance on this dataset.

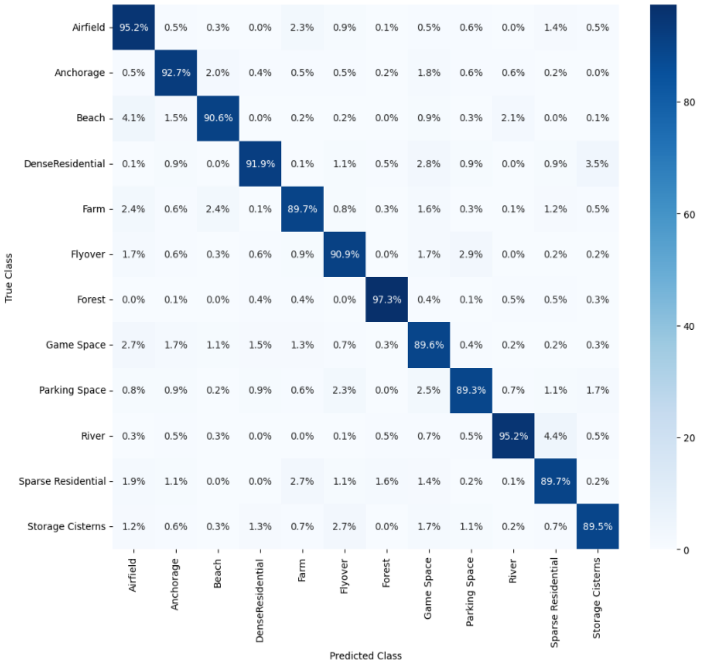


Figure 10: Confusion matrix of WNN classifier using NWPU-RESISC dataset

Table 3: Classification results of proposed M^{2}IAN model on NWPU-RESISC dataset features

Classifier	Accuracy (%)	Recall (%)	Precision (%)	FNR	Time (sec)	Pred Speed (obs/sec)	F1-Score (%)
NNN	84.3	84.26	84.32	15.7	919.02	8600	84.29
MNN	88.7	88.69	88.725	12.3	376.48	8600	88.70
WNN	**91.8**	91.62	91.57	8.20	422.26	8300	91.59

BNN	89.2	89.18	89.17	10.80	926.07	8400	89.17
TNN	91.4	91.36	91.375	22.85	952.09	8700	91.36

B. Discussion

A detailed discussion of the proposed model is presented in this section based on ablation studies, comparisons, generalizability, and interpretation through the XAI technique LIME. Figure 4 shows the architecture of the proposed high-resolution image generation, and output images are shown in Figure 5. The proposed classification architecture is illustrated in Figure 6, and results are presented in Tables 2-3. Moreover, the confusion matrices of both datasets are presented in Figures 8 and 10. Based on the results, it is observed that the proposed model obtained improved performance. However, it is important to analyze the performance of proposed models and their impact on the accuracy; therefore, we performed several ablation studies and comparisons.

Ablation Study 1: The first ablation study conducted a comprehensive analysis of different model configurations for the Mixed Coastal and NWPHU RESISC_45 datasets. Table 4 presents the analysis of this ablation study. The experiment inspects three main configurations: a CNN backbone, a 3rd stage IAM backbone, and a 4th stage IAM backbone, tested on original, generated, and combined datasets. The experiments show that the integration of IAM significantly improves the model's performance, with the IAM of Stage 4 consistently outperforming the IAM of Stage 3 in both datasets. In mixed coastal datasets, the CNN base model gained 82.96 % validation accuracy with the original data but increased to 89.34% when trained both with the original data and with the generated data, signifying that synthetic data positively impacts learning. The addition of Stage 3 IAM improved the validation accuracy to 90.39% validation accuracy, demonstrating the importance of attention modules to focus on the most relevant information. Stage 4 of the IAM refined the model to a maximum validation accuracy of 91.61 %, indicating that the deeper integration of attention mechanisms improved feature learning. However, training time has increased considerably, with the highest configuration taking 4 hrs., and 39 minutes, highlighting the balance between accuracy and computational costs.

In the NWPHU RESISC_45 dataset, a similar trend appears, with the backbone CNN yielding 83.96% validation accuracy from the original dataset, while adding the third phase of the IAM improves the accuracy to 87.6% and the fourth phase of the IAM reaching 91.55% using both the original data and the generated data. This highlights that the impact of the proposed IAM is consistent in different datasets. However, the training time increased in Stage 4 IAM, reaching 4 hrs and 19 minutes. The study shows that the proposed IAM module increases model performance, mainly when used in the generated data. Improvements in accuracy of 91.81% indicate that the model is more reliable and less likely to be classified wrongly. However, computational costs are a vital drawback, especially in real-time applications. In real scenarios such as coastal surveillance and disaster prediction using remote sensing images, using such models on real-time monitoring systems will require optimization strategies like pruning, quantification, and knowledge condensation to reduce inference time.

Table 4: Ablation study on different configurations of models for evaluation

Mixed Coastal						
Model Configurations	Original Data	Generated Data	Original +Generated Data	V(Accuracy)	Precision	Training Time
Backbone CNN	✓			82.96	81.41	1hrs 26m
		✓		89.24	79.75	1hrs 47m
			✓	89.34	88.00	1hrs 59m
Backbone+Stag3 IAM	✓			84.15	82.94	2hrs 36m
		✓		85.04	83.46	2hrs 18m
			✓	90.39	89.91	2hrs 47m
Backbone+Stag4 IAM	✓			90.17	90.07	3hrs 41m
		✓		89.00	89.14	3hrs 29m
			✓	**91.61**	**91.81**	4hrs 39m
NWPHU RESISC_45						
Backbone CNN	✓			83.96	82.74	1hrs 19m
		✓		84.00	83.64	1hrs 38m
			✓	84.65	83.48	1hrs 41m
Backbone+Stag3 IAM	✓			86.15	85.69	2hrs 19m
		✓		86.21	85.94	2hrs 24m
			✓	87.6	86.76	2hrs 31m
Backbone+Stag4 IAM	✓			90.41	89.92	3hrs 35m
		✓		90.05	89.96	3hrs 39m

			✓	**91.45**	**90.98**	4hrs 19m

In the next phase, Table 5 compares the proposed M^{2}IAN model with state-of-the-art models. The comparison is conducted based on parameters, model size, training time, inference time, and validation accuracy. The pre-trained EfficientNetB0 and MobileNetV2 are ideal for real-time applications but have a light architecture and a fast inference, while their validation accuracy is slightly lower. The darkNet53 and ResNet101 are powerful but have high Training costs and long inference times, which make them difficult to deploy in real time. The ResNet50 has gained a high validation accuracy of 91.1% but requires a lot of resources. Compared to the proposed M^{2}IAN model, it is the best choice based on performance, which is 91.6%, size is 24.4MB, and inference time is 3.19 sec, making it very suitable for real-world applications such as remote sensing. Disparate heavy networks that require high-end hardware, M^{2}IAN delivers superior performance without extreme computing above and is ideal for cloud and edge deployment.

Table 5: Comparison of proposed model with the state of the art pre-trained models

Models	**Parameters**	**Model Size**	**Training Time (hrs.)**	**Inference Time (sec)**	**Validation (Accuracy)**
EfficientNetb0	5.3 M	20 MB	4hr 21m	1.54	0.902
DarkNet19	20.8 M	80 MB	5hr 48m	3.46	0.891
DarkNet53	41.6 M	159 MB	8hr 16m	11.17	0.908
ResNet18	11.7 M	14 MB	5hr 11m	3.15	0.897
ResNet50	25.6 M	98 MB	7hr 24m	5.45	0.911
ResNet101	44.6 M	171 MB	9hr 21m	16.53	0.891
Inception V3	23.9 M	91 MB	6hr 35m	4.12	0.903
DenseNet201	20.0 M	77 MB	5hr 19m	4.17	0.908
MobileNetV2	3.5 M	14 MB	3hr 17m	2.51	0.893
Proposed M^{2}IAN	6.9 M	24.4 MB	4hr 39m	3.19	0.916

The proposed model is evaluated with and without color features (CF) in the third ablation study, as described in Table 6. As mentioned in this table, without CF, the model achieved F1 score values of 88.71% and 89.81%, respectively. However, with the use of color features, the F1-Score value is improved to 90.12% and 90.41%, highlighting the effectiveness of the integration of color features as a second input to the proposed network, which improves performance and the consistent increase in performance in both datasets indicates generality.

Table 6: Proposed model evaluation using feeding color features and without feeding color features in the proposed network

Mixed Coastal			
Model	**Without CF**	**With CF**	**F1-score**
Proposed M^{2}IAN	✓		88.71
		✓	90.12
NWPHU RESISC_45			
Proposed M^{2}IAN	✓		89.81
		✓	90.41

1) *Model Generalizability*

To check the generalizability of the proposed model, we selected a large remote sensing dataset such as MLRSNet [46] and performed training and testing. In the training phase, we used the same algorithm (Redfox optimization) to select hyperparameters and then trained on 70% data. The validation accuracy of the proposed model on this dataset was achieved by 92.3%, which was further utilized in the testing phase. In the testing phase, features are extracted, and NN classifiers are employed to ensure classification accuracy. Figure 11 shows the proposed model's testing output on this dataset, which obtained the highest accuracy of 90.6% on the NNN classifier, which is improved than some existing techniques such as [46]. The other classifiers also obtained notable performance, such as 88.5, 89.2, 88, and 85.4%, respectively. In addition, a confusion matrix is plotted in Figure 12, showing how the proposed model effectively classifies each class. This figure also noted that there is a big room for improvement in the proposed model's accuracy and precision rate.

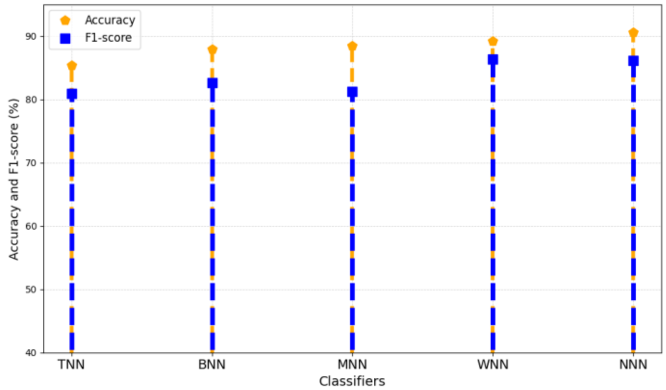


Figure 11: Classification accuracy of proposed model on MLRSNet dataset

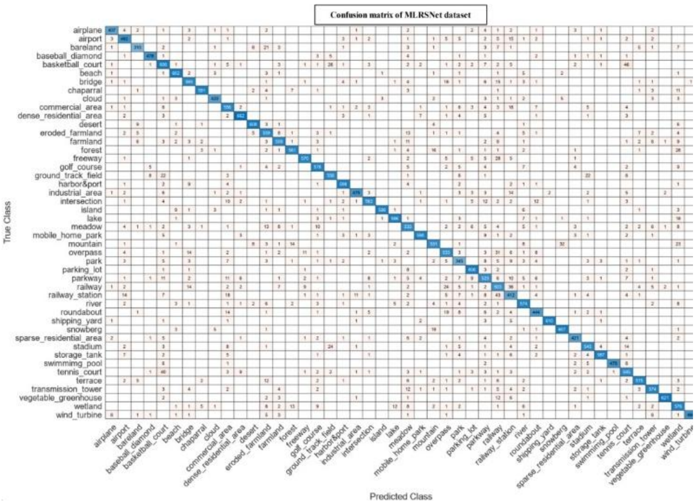


Figure 12: Confusion matrix of NNN classifier for proposed model using MLRSNet dataset

To further interpret the proposed model, we employed an explainable AI technique such as LIME [47]. Figure 13 presents the output of this model on Sampel RS images. In this figure, we added input image, actual label, predicted label correct or incorrect, and LIME interpretation. A few images in this figure are not predicted correctly, like an image collected from the water class, but based on the LIME, it is wrongly predicted into an Island. Hence, it is concluded that the overall proposed model performed well, but there is still room for improving accuracy and precision rates.

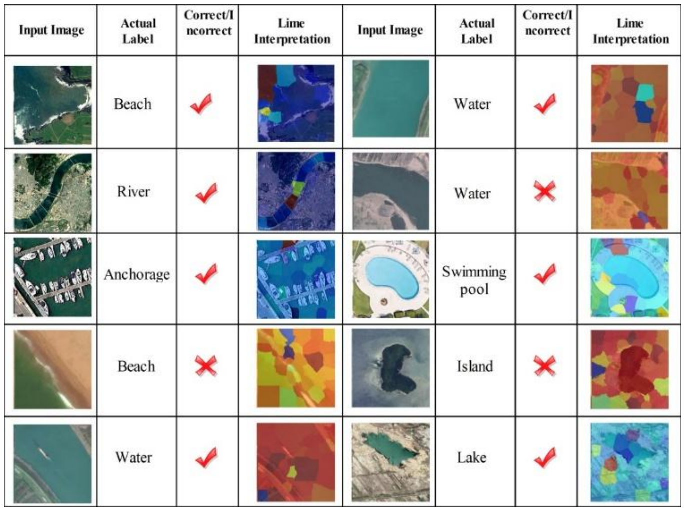


Figure 13: LIME based proposed model interpretation results on RS images

2) *Noise Sensitivity Analysis*

Figure 14 demonstrates the proposed model sensitivity across the noisy data. The synthetic noise has been added in the selected dataset with the ratio of 1.0, 1.3, 1.5, and 2.0%. On Mixed Coastal dataset, the proposed model is showing it is providing high classification accuracy, 90.12% for noise levels 1.0, 1.3 and 1.5; then 89% accuracy for noise level 2.0; this supports the idea that the model is robust when dealing with moderate noise at the coastal image data. For the MLRSNet dataset, the proposed model is achieved accuracy of 90.00% a noise level of 1.0. For NWPHU, performance appears stable at 90.9% for noise levels 1.0 and 1.3; followed by small reductions to 89% at 1.5 where there seems to be a more noticeable reduction in robustness at noise level 2.0(86%). Overall results demonstrated that the proposed architecture is robust to light and moderate noise.

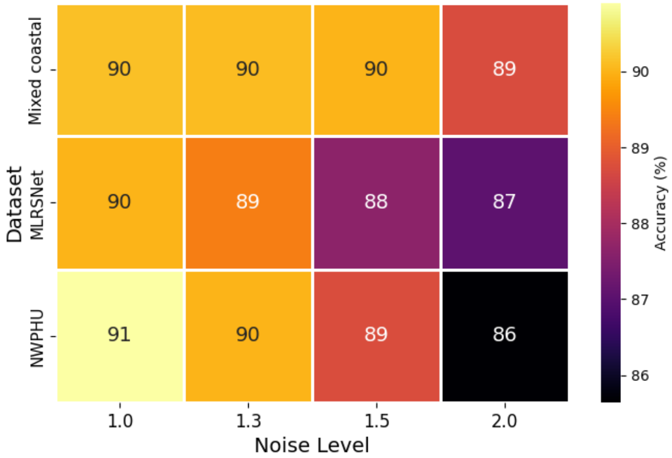


Figure 14: comparison of proposed model by adding the synthetic noise in the datasets

3) *Impact of Transition Layers*

In Figure 15, the impact of transition layers in the proposed model has been evaluated. The figure presented the progress of the proposed M2IAN model accuracy through four distinguishable stages of transition across a range of training epochs up to 70. Each stage of transition represents a step of architectural refinement to the respective model. Each of the respective stages shows an initial trajectory of increased accuracy as the number of training epochs increases, which indicates it is learning effectively across the transition stages. Transition Stage 4 has the final accuracy of 91.8%, representing the most refined level of accuracy. Transition Stage 3 has an accuracy of 88.5%, while Stage 2 has an accuracy of 85.8%, and Stage 1 has an accuracy of 80.7%. After epoch 30 there is a noticeable separation in accuracy between the stages, showing the added architectural improvements of later transition stages.

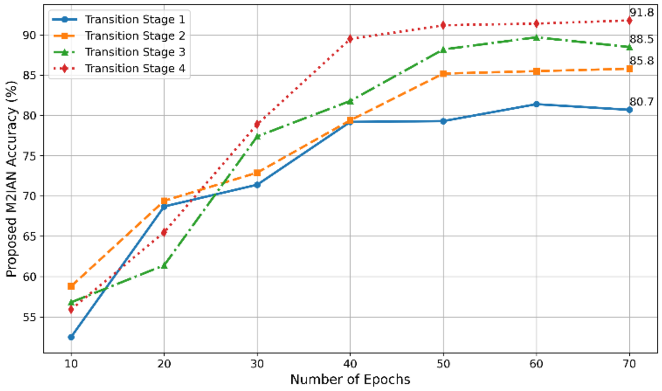


Figure 15: Proposed model performance of transition stages across epochs

In Table 7, a comprehensive comparison is conducted with the existing techniques. In 2011, [48] used an unsupervised learning approach for land-use classification of the Negev region with Landsat-5 TM data. They combined the unsupervised classification with a GIS component to create a hybrid method and obtained an overall accuracy of 81 percent. In 2017, [49] applied the Maximum Likelihood Classifier to detect changes in coastal land cover and land use along the Kanyakumari coast, India, with multi-temporal Landsat imagery and the use of ground truth to assess accuracy, yielding an overall accuracy of 81.16 %. Adding to the comparison is a more recent study by [50], which employed pre-trained deep learning models including UNet++ for land use and crops classification, with Sentinel-2A and Gaofen imagery. Their approach included methods such as feature selection and up sampling to account for class imbalance, and they achieved relatively lower accuracy of 75.34 % likely due to variability in image resolution and resolution among products. In [51] developed a CNN model to classify land covers and crop types in Ukraine with the use of multi-source and multi-temporal satellite data. The deep learning framework contained two layers (one supervised layer and an unsupervised layer), and achieved accuracy values of 85 %, showing a significant improvement from datasets that used traditional classifiers such as random forests and multi-layer perceptron. In comparison, the proposed method that utilized data from multimodal fusion method showed a relatively higher performance, with classification accuracies of 91.8% and 90.6%. The study incorporated multiple modalities of data including images and color features, increasing accurate classification by using the complementary nature of each modality.

Table 7: Comparison between the SOTA and proposed framework

Ref	Year	Methodology	Achieved Results
[51]	2017	Customized CNN	85%
[50]	2022	Pre-trained DL models	75.34%
[48]	2011	Unsupervised learning method	81%
[49]	2017	Detection algorithm	81.16%
Proposed method		MultiModal Fusion	91.8
		-	90.6

V. Conclusion

A novel end-to-end deep learning framework is presented in this work for the image super-resolution and LULC classification from remote sensing images. The proposed framework is based on two novel architectures, SR-RAN5 and M2IAN. The proposed SR-RAN5 is employed to generate high-resolution images that consider the challenge of low-contrast dataset samples. The generated images are replaced and added to the dataset for training the proposed M2IAN architecture. The proposed M2IAN architecture is based on the inception attention modules further fused with color features. The color features impact the classification accuracy due to high-resolution images. In the training of the proposed model, hyperparameters are selected through the Redfox optimization algorithm. The trained model is later utilized for feature extraction and performed classification. Extensive experiments are conducted on three publically available aerial scene datasets: NWPU, MLRSNet, and Mixed Coastal. On these datasets, the proposed model obtained improved accuracy of 91.8, 90.6, and 90.0%, respectively. Based on the obtained results and comparisons, we conclude the following points:

- Generated high-resolution images through SR-RAN5 architecture improved the accuracy of training and testing. A 3-4% improvement occurred in the accuracy after employing a high-resolution generated dataset.
- The proposed M2IAN model and color features improved the classification accuracy. Without color features, there is a decline in the accuracy of almost 4% to 5%. In addition, the proposed model contains less number of parameters and can be employed for real-time RS applications.
- Interpreting the proposed M2IAN architecture was also performed, and we obtained correct predictions and localized important regions.

The proposed has imperfect performance on classes with highly coinciding spectral features and the reliance on manually extracted color features that may not generalize across other sensor modalities. In the future, the proposed model can be implemented on more high-dimensional datasets. In addition, the model can be further fused with ViT blocks to generate better accuracy on MLRSNet datasets. In addition, more interpretation techniques, such as GradCAM and Occlusion Sensitivity, will be implemented.

CONFLICT OF INTEREST

All authors declared no conflict of interest.

DATASET AVAILABILITY

The datasets of this work are publically available for the research purposes. The hyperlinks are; The collected datasets are publically available at (https://1drv.ms/u/s!AnoH-ohvSrcVbimm4dCSqGBnTS4?e=0fu0xO).
The NWPU dataset is publicly available at (https://figshare.com/articles/dataset/NWPU-RESISC45_Dataset_with_12_classes/16674166?file=30871912). MLSRNet Dataset: (https://paperswithcode.com/dataset/mlrsnet).

XI. REFERENCES

[1] F. Liu *et al.*, "Remoteclip: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[2] S. Puttinaovarat, K. Khaimook, and P. Horkaew, "Land use and land cover classification from satellite images based on ensemble machine learning and crowdsourcing data verification," *International Journal of Cartography,* vol. 11, no. 1, pp. 3-23, 2025.

[3] F. Ullah, B. Zhang, R. U. Khan, I. Ullah, A. Khan, and A. M. Qamar, "Visual-based items recommendation using deep neural network," in *Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things*, 2020, pp. 122-126.

[4] T. Qi *et al.*, "Land intensification use scenarios based on urban land suitability assessment of the national park," *Sustainable Cities and Society,* vol. 102, p. 105229, 2024.

[5] S. Das and R. Sarkar, "Role of Anthropogenic Activities on the Riverbank Instability: A Geospatial Analysis on Part of the Bhagirathi-Hugli River, India," in *Contemporary Social Physics: Decoding Social Behaviour with Advanced Geospatial Tools*: Springer, 2025, pp. 217-248.

[6] X. Duan *et al.*, "A geospatial and statistical analysis of land surface temperature in response to land use land cover changes and urban heat island dynamics," *Scientific Reports,* vol. 15, no. 1, p. 4943, 2025.

[7] S. H. Kadhim, S. M. Al-Jawari, and N. A. Razak Hasach, "Analyzing Earth's Surface Temperatures with Relationship to Land Urban Land Cover (LULC) to Enhance Sustainability," *International Journal of Sustainable Development & Planning,* vol. 19, no. 1, 2024.

[8] X. ZHANG, Q. SHI, Y. SUN, J. HUANG, and D. HE, "The Review of Land Use/Land Cover Mapping AI Methodology and Application in the Era of Remote Sensing Big Data," *Journal of Geodesy & Geoinformation Science,* vol. 7, no. 3, 2024.

[9] F. Ullah, B. Zhang, G. Zou, I. Ullah, and A. M. Qamar, "Large-scale Distributive Matrix Collaborative Filtering for Recommender System," in *Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things*, 2020, pp. 55-59.

[10] M. S. Anand, R. A. A. Rosaline, G. Padmapriya, P. Samuel, and P. Kirubanantham, "Develop a Hybrid Ensemble Transfer-Based Residual Multi-Resolution CNN for Classification of Land Cover in Remote Sensing Images," in *Harnessing AI in Geospatial Technology for Environmental Monitoring and Management*: IGI Global Scientific Publishing, 2025, pp. 99-124.

[11] F. Ullah, I. Ullah, K. Khan, S. Khan, and F. Amin, "Advances in deep neural network-based hyperspectral image classification and feature learning with limited samples: a survey," *Applied Intelligence,* vol. 55, no. 6, pp. 1-48, 2025.

[12] T. Yun *et al.*, "Status, advancements and prospects of deep learning methods applied in forest studies," *International Journal of Applied Earth Observation and Geoinformation,* vol. 131, p. 103938, 2024.

[13] S. Mei, J. Lian, X. Wang, Y. Su, M. Ma, and L.-P. Chau, "A comprehensive study on the robustness of deep learning-based image classification and object detection in remote sensing: Surveying and benchmarking," *Journal of Remote Sensing,* vol. 4, p. 0219, 2024.

[14] N. Li, Z. Wang, and F. A. Cheikh, "Discriminating spectral–spatial feature extraction for hyperspectral image classification: a review," *Sensors,* vol. 24, no. 10, p. 2987, 2024.

[15] E. Akdemir and N. Barışçı, "A review on deep learning applications with semantics," *Expert Systems with Applications*, p. 124029, 2024.

[16] Z. Li, B. Chen, S. Wu, M. Su, J. M. Chen, and B. Xu, "Deep learning for urban land use category classification: A review and experimental assessment," *Remote Sensing of Environment,* vol. 311, p. 114290, 2024.

[17] S. S. Sohail *et al.*, "Advancing 3D point cloud understanding through deep transfer learning: A comprehensive survey," *Information Fusion*, p. 102601, 2024.

[18] S. Chakraborty and L. Dey, "Deep Learning-Inspired Multiclass and Multi-label Classifications," in *Multi-objective, Multi-class and Multi-label Data Classification with Class Imbalance: Theory and Practices*: Springer, 2024, pp. 105-134.

[19] L. Alzubaidi *et al.*, "Comprehensive review of deep learning in orthopaedics: Applications, challenges, trustworthiness, and fusion," *Artificial Intelligence in Medicine,* p. 102935, 2024.

[20] X. Zhang, Z. Huang, X. Yao, X. Feng, G. Cheng, and J. Han, "Cross-Modality Domain Adaptation Based on Semantic Graph Learning: From Optical to SAR Images," *IEEE Transactions on Geoscience and Remote Sensing,* 2025.

[21] H. Hichri, "NWPU-RESISC45 Dataset with 12 classes," ed: figshare, 2021.

[22] Z. Zafar, M. Zubair, Y. Zha, S. Fahd, and A. A. Nadeem, "Performance assessment of machine learning algorithms for mapping of land use/land cover using remote sensing data," *The Egyptian Journal of Remote Sensing and Space Sciences,* vol. 27, no. 2, pp. 216-226, 2024.

[23] W. Wang, Z. Guo, Z. Cui, H. Zhao, and L. Xian, "Centripetal Intensive Deep Hashing for Remote Sensing Image Retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 2025.

[24] M. Fayaz, J. Nam, L. M. Dang, H.-K. Song, and H. Moon, "Land-cover classification using deep learning with high-resolution remote-sensing imagery," *Applied Sciences,* vol. 14, no. 5, p. 1844, 2024.

[25] H. M. Albarakati *et al.*, "A novel deep learning architecture for agriculture land cover and land use classification from remote sensing images based on network-level fusion of self-attention architecture," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 2024.

[26] M. Aljebreen, H. A. Mengash, M. Alamgeer, S. S. Alotaibi, A. S. Salama, and M. A. Hamza, "Land use and land cover classification using river formation dynamics algorithm with deep learning on remote sensing images," *IEEE Access,* vol. 12, pp. 11147-11156, 2024.

[27] Y. Liu, Y. Zhong, S. Shi, and L. Zhang, "Scale-aware deep reinforcement learning for high resolution remote sensing imagery classification," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 209, pp. 296-311, 2024.

[28] V. Vinaykumar, J. A. Babu, and J. Frnda, "Optimal guidance whale optimization algorithm and hybrid deep learning networks for land use land cover classification," *EURASIP Journal on Advances in Signal Processing,* vol. 2023, no. 1, p. 13, 2023.

[29] H. Xie, Y. Chen, and P. Ghamisi, "Remote sensing image scene classification via label augmentation and intra-class constraint," *Remote Sensing,* vol. 13, no. 13, p. 2566, 2021.

[30] K. Xu, H. Huang, P. Deng, and Y. Li, "Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 33, no. 10, pp. 5751-5765, 2021.

[31] F. Rauf *et al.*, "FMANet: Super Resolution Inverted Bottleneck Fused Self-Attention Architecture for Remote Sensing Satellite

Image Recognition," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 2024.

[32] Z. Situ *et al.*, "Attention-based deep learning framework for urban flood damage and risk assessment with improved flood prediction and land use segmentation," *International Journal of Disaster Risk Reduction,* vol. 116, p. 105165, 2025.

[33] M. K. Bhatti *et al.*, "A Novel Approach for High-Resolution Coastal Areas and Land Use Recognition from Remote Sensing Images based on Multimodal Network-Level Fusion of SRAN3 and Lightweight Four Encoders ViT," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 2025.

[34] F. Ullah *et al.*, "Deep hyperspectral shots: Deep snap smooth wavelet convolutional neural network shots ensemble for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 17, pp. 14-34, 2023.

[35] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *ACM Computing Surveys (CSUR),* vol. 53, no. 3, pp. 1-34, 2020.

[36] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops,* 2017, pp. 136-144.

[37] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence,* vol. 38, no. 2, pp. 295-307, 2015.

[38] P. Xu, Q. Liu, H. Bao, R. Zhang, L. Gu, and G. Wang, "FDSR: An Interpretable Frequency Division Stepwise Process Based Single-Image Super-Resolution Network," *IEEE Transactions on Image Processing,* 2024.

[39] W. Zhang *et al.*, "CVANet: Cascaded visual attention network for single image super-resolution," *Neural Networks,* vol. 170, pp. 622-634, 2024.

[40] S. Y. Chaganti, I. Nanda, K. R. Pandi, T. G. Prudhvith, and N. Kumar, "Image Classification using SVM and CNN," in *2020 International conference on computer science, engineering and applications (ICCSEA),* 2020: IEEE, pp. 1-5.

[41] C. Sitaula and T. B. Shahi, "Multi-channel CNN to classify nepali covid-19 related tweets using hybrid features," *Journal of Ambient Intelligence and Humanized Computing,* vol. 15, no. 3, pp. 2047-2056, 2024.

[42] S. Kajkamhaeng and C. Chantrapornchai, "SE-SqueezeNet: SqueezeNet extension with squeeze-and-excitation block," *International Journal of Computational Science and Engineering,* vol. 24, no. 2, pp. 185-199, 2021.

[43] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV),* 2018, pp. 3-19.

[44] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *The Visual Computer,* vol. 38, no. 8, pp. 2939-2970, 2022.

[45] D. Połap and M. Woźniak, "Red fox optimization algorithm," *Expert Systems with Applications,* vol. 166, p. 114107, 2021.

[46] X. Qi *et al.*, "MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 169, pp. 337-350, 2020.

[47] M. Henninger and C. Strobl, "Interpreting machine learning predictions with LIME and Shapley values: theoretical insights, challenges, and meaningful interpretations," *Behaviormetrika*, pp. 1-31, 2024.

[48] O. Rozenstein and A. Karnieli, "Comparison of methods for land-use classification incorporating remote sensing and GIS inputs," *Applied Geography,* vol. 31, no. 2, pp. 533-544, 2011.

[49] S. Kaliraj, N. Chandrasekar, K. Ramachandran, Y. Srinivas, and S. Saravanan, "Coastal landuse and land cover change and transformations of Kanyakumari coast, India using remote sensing and GIS," *The Egyptian Journal of Remote Sensing and Space Science,* vol. 20, no. 2, pp. 169-185, 2017.

[50] L. Wang, J. Wang, Z. Liu, J. Zhu, and F. Qin, "Evaluation of a deep-learning model for multispectral remote sensing of land use and crop classification," *The Crop Journal,* vol. 10, no. 5, pp. 1435-1451, 2022.

[51] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters,* vol. 14, no. 5, pp. 778-782, 2017.

Authors Biography

Muhammad Attique Khan (Member IEEE) received the master's and Ph.D. degrees in human activity recognition for application of video surveillance and skin lesion classification using deep learning from COMSATS University Islamabad, Islamabad, Pakistan, in 2018 and 2022, respectively. He is currently an Assistant Professor with AI Department, Prince Mohammad Bin Fahd, Al-Khobar, Saudi Arabia. His primary research focus in recent years is medical imaging, COVID-19, MRI analysis, video surveillance, human gait recognition, and agriculture plants using deep learning. He has above 350 publications that have more than 16 000+ citations and an impact factor of 1050+ with h-index 74 and i-index 230. He is the Reviewer of several reputed journals, such as the IEEE Transaction on Industrial Informatics, IEEE Transaction of Neural Networks, Pattern Recognition Letters, Multimedia Tools and Application, Computers and Electronics in Agriculture, IET Image Processing, Biomedical Signal Processing Control, IET Computer Vision, EURASIP Journal of Image and Video Processing, IEEE Access, MDPI Sensors, MDPI Electronics, MDPI Applied Sciences, MDPI Diagnostics, and MDPI Cancers.

Ameer Hamza is currently working toward the Ph.D. degree in computer science with KTU University, Kaunas, Lithuania. Hamza completed his master degree in computer science from HITEC University in year 2023. His major interests include object detection and recognition, video surveillance, medical, and agriculture using deep learning and machine learning. He has published 20 impact factor papers to date.

Wardah Ibrar completed her bachelor and master degree in computer engineering and computer science from HITEC University, Pakistan in year 2022 and 2024, respectively. Currently she is working at Center of AI, Prince Mohammad bin Fahd University (remotely under Dr. Attique). Her research interest includes remote sensing and medical imaging applications using AI techniques.

Leila Jamel received the engineering degree in computer sciences and the Ph.D. degree in computer sciences and information systems from the National School of Computer Sciences, Tunisia, in 2008.,She is currently an Assistant Professor with the Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University (PNU), Riyadh, Saudi Arabia. She is a Researcher with RIADI Laboratory-Tunisia. Her research interests include business process reengineering, process modeling, BPM, data sciences, ML, process mining, e-learning, and software engineering.,Dr. Jamel was the Program Leader of the IS program and the ABET and NCAAA accreditation committees with the CCIS, PNU. She worked as a Head of the Information Systems Security Department with the Premier Ministry in Tunisia. She is a Reviewer of many international journals and conferences. She was a Member of scientific/steering committees of many international conferences.

Mehrez Marzougui was born in Kasserine, Tunisia, in 1972. He received the B.Sc. degree from the University of Tunis, Tunis, Tunisia, in 1996, and the M.Sc. and Ph.D. degrees from the University of Monastir, Monastir, Tunisia, in 1998 and 2005, respectively, all in electronics. From 2001 to 2005, he was a Research Assistant with Electronics and Micro-Electronics Laboratory. From 2006 to 2012, he was an Assistant Professor with Electronics Department, University of Monastir. Since 2013, he has been an Assistant Professor with Engineering Department, College of Computer Science, King Khalid University, Abha, Saudi Arabia. He is the author of more than 30 articles. His research interests include hardware/software cosimulation, image processing, and multiprocessor systems on chips.

Saru Kumari received the Ph.D. degree in mathematics in 2012 from Chaudhary Charan Singh (CCS) University, Meerut, India.,She is currently an Assistant Professor with the Department of Mathematics, CCS University, Meerut, India. Her research interests include information security and security of WSN.

Areej Alasiry received the B.Sc. degree in information systems from King Khalid University, Abha, Saudi Arabia, and the M.Sc. degree (Hons.) in advanced information systems and the Ph.D. degree in computer science and information systems from Birkbeck College, University of London, U.K., in 2010 and 2015, respectively. She is currently an Assistant Professor at the College of Computer Science, King Khalid University. She also holds the position of the College Vice Dean for Graduate Studies and Scientific Research. Her main research interests include machine learning and data science.

Yunyoung Nam (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer engineering from Ajou University, South Korea, in 2001, 2003, and 2007, respectively. He was a Senior Researcher with the Center of Excellence in Ubiquitous System, Stony Brook University, Stony Brook, NY, USA, from 2007 to 2010, where he was a Postdoctoral Researcher, from 2009 to 2013. He was a Research Professor with Ajou University, from 2010 to 2011. He was a Postdoctoral Fellow with Worcester Polytechnic Institute, Worcester, MA, USA, from 2013 to 2014. He was the Director of the ICT Convergence Rehabilitation Engineering Research Center, Soonchunhyang University, from 2017 to 2020. He has been the Director of the ICT Convergence Research Center, Soonchunhyang University, since 2020, where he is currently an Assistant Professor with the Department of Computer Science and Engineering. His research interests include multimedia database, ubiquitous computing, image processing, pattern recognition, context-awareness, conflict resolution, wearable computing, intelligent video surveillance, cloud computing, biomedical signal processing, rehabilitation, and healthcare systems.